From multilingual web-archives to parallel treebanks in five minutes

From Multilingual Web-Archives to Parallel Treebanks in Five Minutes

Markus Killer, Rico Sennrich, Martin Volk
University of Zurich
Institute of Computational Linguistics, Binzmühlestrasse 14, CH-8050 Zurich, Switzerland
E-mail: markus.killer@uzh.ch, sennrich@cl.uzh.ch, volk@cl.uzh.ch

Abstract

The Tree-to-Tree (t2t) Alignment Pipe is a collection of Python scripts, generating automatically aligned parallel treebanks from multilingual web resources or existing parallel corpora. The pipe contains wrappers for a number of freely available NLP software programs. Once these third party programs have been installed and the system and corpus specific details have been updated, the pipe is designed to generate a parallel treebank with a single program call from a unix command line. We discuss alignment quality on a fully automatically processed parallel corpus.

Keywords: parallel treebank, automatic tree-to-tree alignment, TreeAligner, Text-und-Berg

1. Introduction

The process of creating parallel treebanks used to be a tedious task, involving a tremendous amount of manual annotation (see e.g. Samuelsson & Volk, 2007). Zhechev and Way (2008:1) state that ”[b]ecause of this, only a few parallel treebanks exist and none are of sufficient size for productive use in any statistical MT application”. Since Zhechev (2009) introduced the *Sub-Tree Aligner*, a program for the automatic generation of parallel treebanks, the feasibility of obtaining large scale annotated parallel treebanks has increased. However, the amount of preprocessing needed as well as the missing conversion of the output into a more human readable format might have kept potential users of the *Sub-Tree Aligner* at a distance. The collection of Python scripts combined in the *Tree-to-Tree Alignment Pipe* (*t2t-pipe*) described below takes care of all necessary pre- and postprocessing of Zhechev’s *Sub-Tree Aligner*, supporting German, French and English as source and target languages. The focus of this paper is on the following two questions, both aimed at maximizing the quality of the automatic alignments:

- How big does the parallel corpus have to be in order to get satisfactory results?
- What can be said about the role of the text domain/topic of the parallel corpus?

2. Related Work

Zhechev (2009) and Koehn (2009) provide an overview of recent developments in tree-to-tree alignment, subtree alignment and the subsequent generation of parallel treebanks for use in statistical machine translation systems.

Tiedemann and Kotzé (2009) and Tiedemann (2010) propose a supervised approach to tree-to-tree alignment, requiring a small manually aligned or manually corrected treebank of at least 100 sentence pairs[1] for training purposes.

In terms of script design, the training-script for the *Moses* SMT system (Koehn, 2010b) inspired the organization of the *t2t-pipe* into several steps that can be run independently.

3. Parallel Corpora

In an ideal world, one could be inclined to take a number of parallel articles from a bilingual text collection and let the *t2t-pipe* combined with the *Sub-Tree Aligner* do the rest. Yet this is only possible if a suitable word alignment model[2] is available, as we will show in section 5 .

[1] See http://stp.lingfil.uu.se/~joerg/Lingua/index.html (accessed: 21/08/11)
[2] All word alignment models used in this paper can be downloaded from: http://t2t-pipe.svn.sourceforge.net/ (accessed: 21/08/11)

With the aim of collecting information on the role of corpus size and text domain/topic in creating an automatically aligned parallel treebank, the following corpora were used:

3.1. Corpus for Tree-to-Tree Alignment

A subcorpus of the Text+Berg corpus (Volk et al., 2010) consisting of four parallel articles from the Swiss Alpine Club Yearbook 1977 served as test corpus (see [TUB-4-ART] in table 1). Details on the corpus with regard to the extraction of parallel articles and sentence pairs are described in Sennrich and Volk (2010). For the purpose of this paper it is sufficient to note that the vast majority of texts can be attributed to the journalistic textual domains article/report/review with a strong topical focus on activities performed by members of the Swiss Alpine Club (climbing, hiking, trekking) and the alpine environment in general. As the corpus has been digitised from printed books it contains OCR errors.

Corpus	Lang.	Tokens	Sentence Pairs
[TUB-4-ART]	DE	21,689	1,171
	FR	25,388	(*GIZA++*: 1,023)
[TUB]	DE	1,617,301	92,518
	FR	1,921,583	(*GIZA++*: 80,698)
[EPARL]	DE	35,371,164	1,562,563
	FR	42,427,755	(*GIZA++*: 1,190,609)

Table 1: Parallel Corpora

[TUB-4-ART] Text+Berg Corpus 4 Articles SAC YB 1977
[TUB] Text+Berg Corpus SAC Yearbooks 1957-1982
[EPARL] Europarl Corpus 1996-2009

3.2. Corpora for Word Alignment

Additionally, we used the complete Text+Berg corpus [TUB] , the Europarl corpus (Koehn, 2010a) [EPARL] and combinations of these two corpora to compute different word alignment models (see table 1 for basic corpus information). Word alignment is automatically computed through *GIZA++* (Och & Ney, 2003), which implements the IBM word alignment models. For performance reasons, we set the maximum sentence length to 40 tokens[3]. Therefore, we used only 83% of the of the [TUB] corpus and 76% of the [EPARL] corpus to estimate word alignment probabilities (see table 1 for absolute values in brackets).

We used [EPARL] to test the impact of corpus size on the results. Moreover, texts from the [EPARL] corpus belong to a completely different textual domain (parliament proceedings) and cover a wide range of political, economic and cultural topics (see Koehn, 2009:53), making it possible to use the data to figure out the role of text domain/topic in the alignment process.

4. The *t2t-pipe*

Taking an existing parallel corpus[4] as input, the *t2t-pipe* runs through seven steps to generate automatic alignments for individual words and syntactic constituents in each parallel sentence pair. The configuration file is deliberately designed in a way that a number of different third party programs can be chosen for most of the steps, enabling easy switching between different configurations. In the brief outline of the following steps, the configuration that worked best is indicated (please refer to the *t2t-pipe README* file[5] for details on all 12 programs used):

4.1. Steps 1-5 – Preprocessing

1) Extraction of Parallel Articles
2) Tokenization
   (*Python NLTK Punkt-Tokenizer*)
   Rudimentary OCR cleaning/
   Fixing of word division errors
3) Sentence Alignment
   (*Hunalign* with *dict.cc* dictionary)
4) Statistical Phrase Structure Parsing
   (*Stanford Parser* for German,
   Berkeley Parser for French)
5) Word Alignment
   (*GIZA++* through *Moses* training script,
   enhanced with *dict.cc* dictionary,
   see section 4.2 for an example),
   data not lower-cased

[3] See http://www.statmt.org/wmt11/baseline.html (accessed: 21/08/11)

[4] If no parallel corpus is available, the pipe includes scripts for the on-the-fly construction of a parallel corpus from the web archives of the bilingual Swiss Alpine Club magazine (German-French).

[5] Available from: http://t2t-pipe.svn.sourceforge.net/ (accessed: 21/08/11)

4.2. Step 6 - Tree-to-Tree Alignment

This is the most important step in a complete run of the *t2t-pipe*, as the automatic alignments are generated by Zechev's *Sub-Tree Aligner*. The process can best be described by looking at a parallel sentence pair, taken from [TUB-4-ART]:

1) German sentence: *Man versuche einmal einen solchen Mann abzubremsen.*
2) French sentence: *Essayez donc de freiner un tel homme.*[6]

- Input:

a. Bracketed parse trees of source and target language (output of the two parsers combined into one file):

```
(ROOT (NUR (S (PIS Man) (VVFIN versuche) (ADV einmal) (VP (NP (ART einen) (PIDAT solchen) (NN Mann)) (VVIZU abzubremsen))) ($. !))) \n
(ROOT (SENT (VN (V Essayez)) (ADV donc) (VPinf (P de) (VN (V freiner)) (NP (D un) (A tel) (N homme))) (. !)))\n\n\n
```

b. Two lexical translation files generated by the *Moses* training script and *GIZA++*, enhanced using a *dict.cc* dictionary:

lex.e2f (French – German – Probability)

```
Homme Mann 1.0000000
homme Mann 1.0000000
mari Mann 1.0000000
ralentir abzubremsen 0.0666667
freiner abzubremsen 0.0666667
```

lex.f2e (German – French – Probability)

```
abzubremsen ralentir 0.0053476
abzubremsen freiner 0.0035842
Mann Homme 1.0000000
Mann homme 1.0000000
Mann mari 1.0000000
```

- Output:

  Indexed bracketed parse trees of source and target language with alignment indices on a separate line (see Figure 1 for graphical alignments). In our example sentence, the *Sub-Tree Aligner* produced one wrong alignment, linking the German personal pronoun *man* to the French finite verb *essayez (emphasised below)*:

```
(ROOT::NUR-2 (S-3 (PIS-4 Man)(VVFIN-5 versuche) (ADV-6 einmal)(VP-7 (NP-8 (ART-9 einen)(PIDAT-10 solchen)(NN-11 Mann))(VVIZU-12 abzubremsen)))($.-13 !)) \n
(ROOT::SENT-2 (VN::V-4 Essayez)(ADV-5 donc) (VPinf-6 (P-7 de)(VN::V-9 freiner)(NP-10 (D-11 un)(A-12 tel)(N-13 homme)))(.-14 !)) \n
2 2 4 4 6 5 7 6 8 10 9 11 10 12 11 13 12 9 13 14
```

4.3. Step 7 - Conversion to *TigerXML*/*TMX*

We converted the output of Zhechev's *Sub-Tree Aligner* into two language specific *TigerXML* files and an additional *XML* file containing information on node alignments. These files can be easily imported into the graphical interface of the *Stockholm TreeAligner* (Lundborg et al., 2007). Figure 1 shows the previously introduced sentence pair – including the automatically computed links – in the treebank browser perspective of the *Stockholm TreeAligner*.

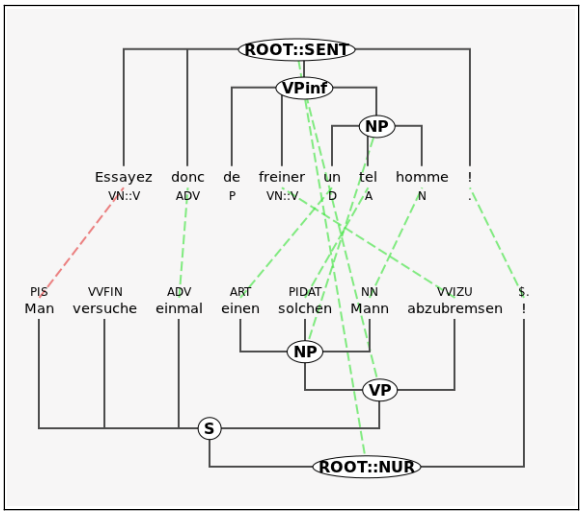

Figure 1: Automatically aligned sentence pair in *Stockholm TreeAligner*

The second supported output format is *TMX*, a format for current translation memory systems (tested with OmegaT[7]).

5. Treebank Alignment Quality

We ran six experiments (summarized in table 2) on the test corpus [TUB-4-ART] (see table 1). In each experiment, the corpus used to compute the lexical translation probabilities with *GIZA++* either differed

[6] Sentences 1) and 2) translate roughly as: [(*Why don't) you try to slow down a man like that (a heavy man)!]*

[7] Available from: http://www.omegat.org (accessed: 21/08/11)

Corpus	**1** [TUB-4-ART]	**2** [TUB-4-ART]	**3** [EPARL]	**4** [TUB]	**5** [TUB-EPARL]	**6** [TUB-EPARL]
Corpus Size *GIZA*++	1,023 SP	1,023 SP	1,190,609 SP	80,698 SP	258,971 SP	1,271,307 SP
In-domain (%)	100.0%	100.0%	0.0%	100.0%	31.0%	6.0%
***Dict.cc* SA/WA**	NO	YES	YES	YES	YES	YES
Precision WA	57.8%	61.1%	51.3%	65.9%	69.1%	69.2%
Precision PhA	58.3%	65.4%	51.8%	81.7%	79.5%	80.4%
Precision allA	57.9%	62.1%	51.4%	69.2%	71.3%	71.7%
Correct links per SP	8.66	9.63	9.02	12.48	13.64	13.98

Table 2: Alignment precision and average number of correct links in treebank of **[TUB-4-ART]** corpus (1,171 sentence pairs) with respect to size, enhancement through additional lexical resources and textual domain of the corpus used to compute the lexical translation probabilities.

Precision = Correct Alignments / Suggested Alignments, SP: Sentence Pair(s) SA: Sentence Alignment, WA: Word Alignment, PhA: Phrase Alignment, allA: Word & Phrase Alignments, In-domain: domain correspondence of treebank and WA corpus

with respect to corpus size and textual domain or enhancement by external lexical resources (*dict.cc* dictionary).

We manually checked an average of 545 alignments (77% word alignments 23% phrase alignments) in 32 randomly selected sentence pairs[8] for each of the six resulting treebanks, using the *Stockholm TreeAligner*. Our information on changes in recall is based on the absolute number of correct links in the manually checked sentence pairs (average no. of correct links = average no. of all links[9] x precision[10]).

5.1. Corpus Size

Looking at the configuration outlined in section 4, three of the seven steps in the *t2t-pipe* directly depend on the corpus size (Tokenization (Dehyphenation), Sentence Alignment and Word Alignment). The analysis of the alignment quality in the resulting parallel treebank shows that roughly 1000 sentence pairs are not enough to get satisfactory results with an overall precision of 57.9% (see table 2, experiment 1). Initial tests have shown that Zhechev's *Sub-Tree Aligner* is highly dependent on the quality of the word alignments supplied. Even though the algorithm does not directly replicate the *GIZA*++ alignments:

> [M]y system uses a probabilistic bilingual dictionary derived from the *GIZA*++ word alignments, thus being able to side-step errors present in the original word-alignment data and to find new possible alignments that *GIZA*++ had skipped for the particular sentence pair.
>
> (Zhechev, 2009:73)

We employed two measures to increase the precision of the alignments:

1) We enhanced the lexical translation probabilities computed by *GIZA*++ by extracting all 1-to-1 word translations from the freely available *dict.cc* dictionary (DE-FR), leading to a substantial increase in precision (+ 4.2%) and in recall (+ 0.97 correct links per sentence pair).
2) Step-by-step, we increased the corpus size, making use of all available resources. In experiment 3 it becomes clear that a huge increase of corpus size alone is no guarantee for better alignment results: When we use the 1,190,609 sentence pair [EPARL] corpus on its own, the recall drops by 0.61 correct

[8] This number proved to be sufficient to include at least 100 Phrase Alignments in the sample. The identity of the treebank was masked for the manual evaluation.
[9] computed by *Sub-Tree Aligner* for the whole treebank
[10] computed from manually checked sentence pairs

links per sentence pair and the precision by 10.7% compared to experiment 2. However, increasing the size of the [TUB] corpus from 1,023 to 80,698 sentence pairs as a basis for the word alignment model leads to the biggest leap in the experiment sequence in both precision (+ 7.1%) and recall (+2.85 correct links per sentence pair) compared to experiment 2.

5.2. Domain/Topic Specific Content

The data collected in table 2 suggests that when using the unsupervised approach proposed by Zhechev (2009) the domain of the corpus used to compute the lexical translation probabilities seems to be of great importance. In experiment 3, we observe the poorest precision of all experiments with the second biggest corpus [EPARL]. Apart from a few common lexical items (e.g. *mountain, valley, river, ...*) there is hardly any overlap in terms of textual domain/topic (see section 3) and the [TUB- 4-ART] corpus itself was not used to compute lexical probabilities in experiment 3 (hence the 0% correspondence between the two corpora).

Comparing these results to the supervised approach by Tiedemann and Kotzé (2009), there seems to be an important difference, as they observe ”only a slight drop in performance when training on a different textual domain” (204) . The main reason for this might be that in the supervised approach the program trains phrase alignments from manually aligned training data (relatively domain/topic independent), whereas in the unsupervised approach the parallel corpus is used to compute lexical translation probabilities (heavily dependent on domain/topic).

5.3. The Right Balance of Corpus Size and Domain/Topic Specific Content

Bearing this difference of the two approaches in mind, it is not surprising that balancing (in terms of textual domain/topic - experiment 5) or expanding (maximising corpus size - experiment 6) the word alignment model affects the results in a different way:

> When using a better model for estimating lexical probabilities (more data: Europarl+SMULTRON) the performance improves only slightly to about 58.64% [F-Score compared to 57.57%]
>
> (Tiedemann & Kotzé, 2009:204)

In the unsupervised approach (used in the *t2t-pipe*) however, the use of a better word alignment model [TUB-EPARL] increases the recall by another 1.16 and 1.50 correct links per sentence pair, respectively (experiments 5/6), compared to the largest corpus with a 100% domain correspondence (experiment 4). For phrase alignments, we achieved a precision of roughly 80% from a corpus size of approx. 80,000 sentence pairs of the same domain (experiments 4-6). The maximum precision of word alignments in this set-up (data not being lower-cased) seems to be around 70% from a corpus size of about 250,000 sentence pairs, while the recall can still be slightly increased by supplying more and more data to estimate lexical probabilities. As long as there is a solid basis of several 10,000 sentence pairs belonging to the same textual domain as the parallel corpus to be aligned, expanding the corpus used to compute lexical probabilities with material of another textual domain does not seem to harm the results but can still help to increase overall precision and recall by a small margin.

6. Conclusion and Outlook

We designed the *t2t-pipe* considering the following areas of application:

1) Assisting human annotators of a parallel treebank by supplying good alignment suggestions: The results discussed in section 5 have shown that this can be achieved by employing a large enough parallel corpus of approx. 250,000 sentence pairs with data of the same textual domain. If the corpus is not big enough, the results can be improved by adding language material of a completely different textual domain. We achieved an overall precision of 71.7% (approx. 80% for phrase alignments). Using a corpus of 500-1,000 sentence pairs (a common size for human annotated parallel treebanks) or a word alignment model trained solely on a different textual domain does not lead to reasonable automatic alignments. However, if there already is a suitable word alignment model for a specific text

domain/topic, the generation of a brand new treebank is just five minutes away.

2) Visualisation/manual evaluation of the results of different components of a tree-based SMT system (e.g. Parsing, Word/Phrase Alignment): The data collected and analysed in section 5 is one possible application of the *t2t-pipe* in this category.
3) As a by-product, the *t2t-pipe* produces phrase alignments for translation memory systems: With a corpus of approx. 80,000 sentence pairs, the precision of the alignments is around 80%. These alignments can be manually checked and a new *TMX* file can be easily generated from the corrected alignment data.

In future versions of the program, the two approaches presented by Zhechev (2009) and Tiedemann and Kotzé (2009) could be combined. We see additional potential for improvement in using lower-cased data and a corpus free of OCR errors for word and subtree alignment.

7. References

Koehn, P. (2009): Statistical Machine Translation. Cambridge: Cambridge University Press.

Koehn, P. (2010a): European Parliament Proceedings Parallel Corpus 1996-2009. Release v5. TXT- Format. Descrition in: Europarl: A Parallel Corpus for Statistical Machine Translation, Philipp Koehn, MT Summit 2005. URL: http://www.statmt.org/europarl.

Koehn, P. (2010b): MOSES. Statistical Machine Translation System. User Manual and Code Guide, November. URL: http://www.statmt.org/moses/manual/manual.pdf.

Lundborg J., Marek T., Mettler M., Volk, M. (2007): Using the Stockholm TreeAligner. In Proceedings of the Sixth International Workshop on Treebanks and Linguistic Theories (TLT'06). Bergen, Norway: Northern European Association for Language Technology, pp. 73–78.

Och, F. J., Ney, H. (2003): A Systematic Comparison of Various Statistical Alignment Models. Computational Linguistics 29, pp. 19–51.

Samuelsson, Y. , Volk, M. (2007): Alignment Tools for Parallel Treebanks. In Proceedings of the GLDV Frühjahrstagung, Tübingen, Germany.

Sennrich R., Volk, M. (2010): MT-based Sentence Alignment for OCR-generated Parallel Texts. In Proceedings of the Ninth Conference of the Association for Machine Translation in the Americas (AMTA 2010).

Tiedemann J., Kotzé, G. (2009): Building a Large Machine-Aligned Parallel Treebank. In Proceedings of the Eighth International Workshop on Treebanks and Linguistic Theories (TLT'08). Milano, Italy: EDUCatt: pp. 197–208.

Tiedemann J. (2010): Lingua-Align: An Experimental Toolbox for Automatic Tree-to-Tree Alignment. In Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'2010), Valetta, Malta.

Volk, M., Bubenhofer, N., Althaus A., Bangerter, M., Marek T., Ruef, B. (2010): Text+Berg-Korpus (Pre-Release 118+ Digitale Edition Die Alpen 1957-1982). XML-Format, May. Digitale Edition des Jahrbuch des SAC 1864-1923 und Die Alpen 1925-1995. URL: http://www.textberg.ch.

Zhechev V., Way, A. (2008): Automatic Generation of Parallel Treebanks. In Proceedings of the 22nd International Conference on Computational Linstuistics. Manchester, UK: pp. 1105–1112.

Zhechev, V. (2009): Automatic Generation of Parallel Treebanks. An Efficient Unsupervised System. Dissertation, School of Computing, Dublin City University.